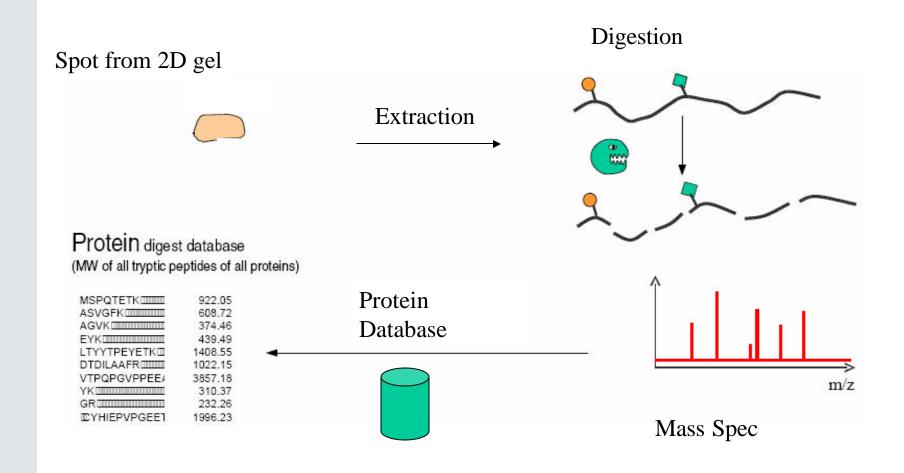


# **Proteomics SIG**

- Spectrometry-based Proteomics
- Standards and Data Exchange



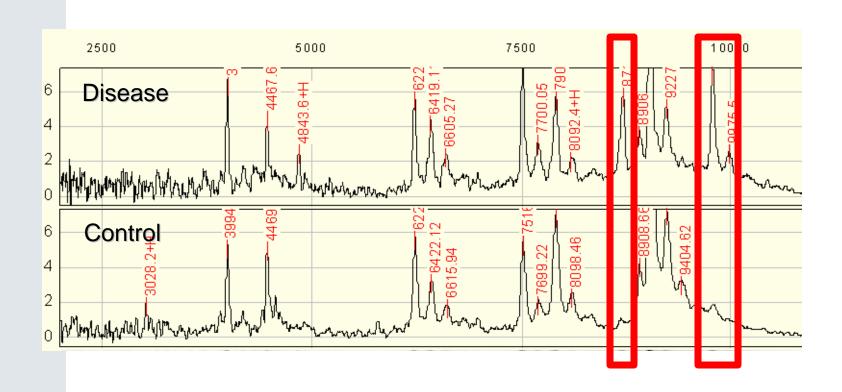
#### **Identification Studies**







# **Profiling Studies**

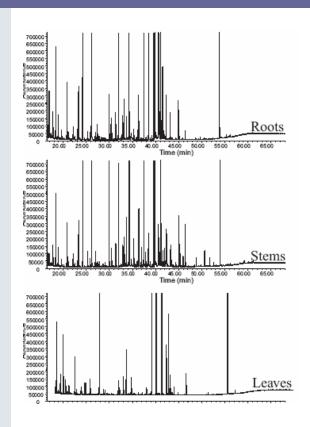


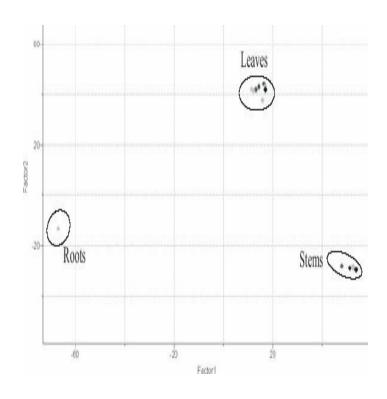
Courtesy: Ciphergen





### **Matabolomics**

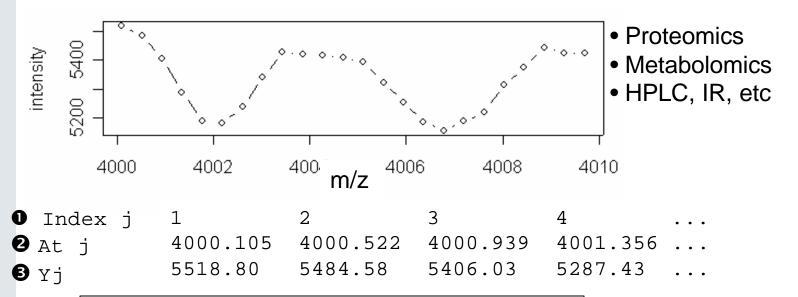






caBIG cancer Biomedical

#### **Spectrum Data**



- 1 Index j: also called clock tick, scan #, sample #, variable #
- 2 At j: also called m/z, mass
- 3 Yj: also called intensity, relative intensity, standardized intensity, abundance





### **Ontology and Common Data Element**

of the optimization methodology, this method can be applied to separations where no prior knowledge of the feed mixture composition exists, which is frequently the case in industrial applications.

#### Nomenclature

A	user-adjustable weight
$a_i$	response model parameters
B	user-adjustable weight
COF	chromatographic optimization function
$f_i$	measure of separation of peak pair i (Figure 2)
$g_i$	measure of separation of peak pair i (Figure 2)
$K_i$	peak geometry penalty for peak i
M	number of expected peaks
N	actual number of eluted peaks
$n_{\rho}$	number of peaks exhibited on chromatogram
$R_{ij}$	resolution between peaks <i>i</i> and <i>j</i>
$\sum R_{ij}$	summation of individual peak pair resolutions over the entire chromatogram
t	total analysis time

- tion of Phenylurea Pesticides using Ternary Mobile Phase Gradients in Reversed-Phase HPLC. *J. Liq. Chromatogr.* 1991, 14, 3125–3151.
- Klein, E.; Rivera, S. Neural Network Signal Interpretation for Optimization of Chromatographic Protein Purifications. Appl. Math. Computer Sci. 1998, 8, 865–886.
- Klein, E.; Rivera, S. A Review of Criteria Functions and Response Surface Methodology for the Optimization of Analytical Scale HPLC Separations. J. Liq. Chromatogr. Relat. Technol. 2000, submitted for publication.
- Lindberg, W.; Johansson, E.; Johansson, K. Application of Statistical Optimization Methods to the Separation of Morphine, Codeine, Noscapine and Papaverine in Reversed-Phase Ion-Pair Chromatography. J. Chromatogr. 1981, 211, 201– 212.
- Lundell, N.; Markides, K. Optimization Strategy for Reversed-Phase Liquid Chromatography of Peptides. J. Chromatogr. 1993, 639, 117–127.
- Palasota, J.; Leonidou, I.; Palasota, J.; Chang, H.-L.; Deming, S. Sequential Simplex Optimization in a Constrained Simplex Mixture Space in Liquid Chromatography. Anal. Chim. Acta 1992, 270, 101–106.
- Wang, Q.-S.; Gao, R.-Y.; Yan, B.-W. Computer-Assisted Optimization of pH and Ion Concentration Selectivity in HPLC

Biotechnol. Prog., 2000, Vol. 16, No. 3





#### **Proteomics: Needs and Standards**

#### Needs

- Tracking lab work flows -- LIMS
- Data storage and retrieval -- Proteomic Databases
- From data to biomarkers Analytical Algorithms
- Interoperation among analytical systems Grid Computing

#### Standards

- MIAPE: Minimum Information About Proteomics Experiments
- SMOS: Statistical Model Of Spectra (under development at Duke)





#### Organism SampleOrigin analyte processing species\_name description MobilePhase Column strain\_identifier condition AssayDataPoint Component relevant\_genotype OntologyEntry description condition\_degree description manufacturer category environment part number **TaggingProcess** concentration protein\_assay value tissue\_type batch number 0.,1 lysis\_buffer description cell\_type internal length PercentX tag\_typ cle phase internal\_diameter 9.1 OtherAnalyte 2...n percentage coolyte parameters • ProcessingStep stationary\_phase ation bead\_size OtherAnalyte name GradientStep metapolic\_tabel pore size ...al\_volume step\_time name temperature AnalyteProcessingStep Experiment flow\_rate Sample, hypothesis Fraction injection volume sample id Analyte Gel method\_citations parameters\_file start\_point sample\_date result\_citations description end\_point experimenter raw\_image ChemicalTreatment protein\_assa Gelltem annotated image ion MassSpecMachine software\_version MassSpecExperiment uderi itisations TreatedAr Vte area manufacturer warped\_image description model name intensity warping\_map parameters\_file Gel<sub>1D</sub> local\_background Band software version equipment denaturing\_agent annotation ercent\_acrylamide tane\_number IonSource mass\_start annotation\_source apparent\_mass solubilization\_buffer Electrospray mass end type. volume stain details spray\_tip\_voltage run\_details collision\_energy pixel\_x\_coord protein assay Spot spray tip diameter pixel\_y\_coord in-gel\_digestion 0..1 apparent\_pi solution\_voltage Gel2D pixel radius background apparent\_mass cone\_voltage normalisation pi\_start pixel\_size\_x loading\_type 0...1 mzAnalysis normalised volume pi\_end pixel\_size\_y BoundaryPoint solvent mass\_start type interface manufacturer pixel\_x\_coord mass end spray\_tip\_manufacturer DiGEGel pixel\_y\_coord first dim details RelatedGelItem dye\_type second dim details description Detection DiGEGelltem excitation\_wavelength gel\_reference exposure\_time dye\_type MSMSFraction item\_reference tiff\_image target\_m\_to\_z )uadrupole matrix\_type plus\_or\_minus DBSearch grid voltage description acceleration\_voltage PeakList username 1... list\_type ion\_mode Tandem id date pole ListProcessing SequenceData n-terminal\_aa description ription c-terminal\_aa Otherionisation mass\_value\_type smoothing\_process source\_type background\_threshold count\_of\_specific\_aa IonTrap sequence name name\_of\_counted\_aa iomiset ion... gas\_type Peak regex\_pattern OntologyEntry gas\_pressure m to z PeptideHit rf\_frequency category abundance excitation\_amplitude value e multiplicity pe description isolation\_centre ProteinHit isolation\_width database all\_peptides\_matched information OthermzAnalysis final\_ms\_level database\_date Chromatogram probability parameter f name CollisionCell Protein ToF gas\_type ccession\_number gas\_pressure reflectron\_state OntologyEntry max missed cleavages gene\_name collision\_offset internal\_length PeakSpecific. category mass value type synonyms ChromatogramIntegration value fragment\_ion\_tolerance organism resolution peptide\_mass\_tolerance description orf number PEDRo UML Class Diagram: Key to colours software version accurate mass mode description Sample Generation Sample Processing RelatedGelItem background\_threshold mass error type sequence area under curve description modifications mass\_error Mass Spectrometry MS Results Analysis peak\_description gel reference predicted\_mass protonated sister\_peak\_reference predicted\_pi icat option item reference

#### **MIAPE**



#### **Statistical Model Of Spectra (SMOS)**

#### Scope

Spectrum-based profiling methods, such as proteomics and metabonomic

#### Focus

Statistical modeling

#### Purpose

- Standard for statistical data analysis, exchange, comparison, and verification
- Audit trail for statistical manipulation of spectral data





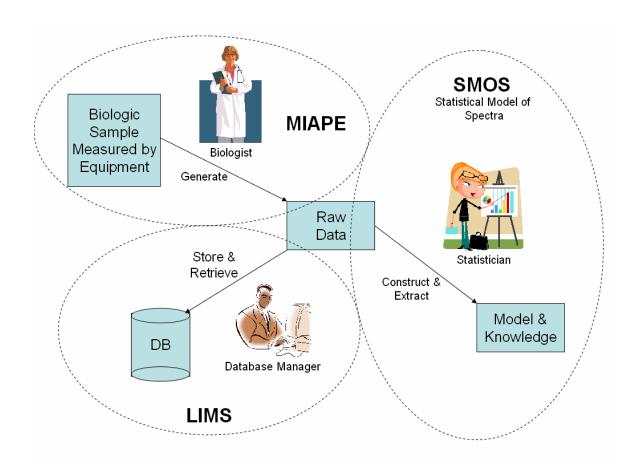
### Statistical Modeling of Spectra (SMOS)

- Single spectrum
  - Baseline removal, Smoothing etc.
- A collection of spectrum
  - Normalization, Aggregation, Alignment etc.
- Raw spectrum -> Extracted Features
  - Peaks, Bins, Principle components
- Extracted Features -> Models
  - Clustering, Classification, and Survival
  - Biomarker discovery



[Next: graphic models of SMOS] cable (Most informatics Grid

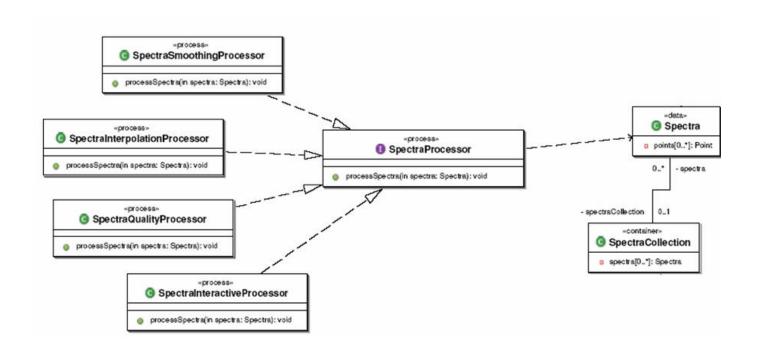
## **Relationship between MIAPE and SMOS**







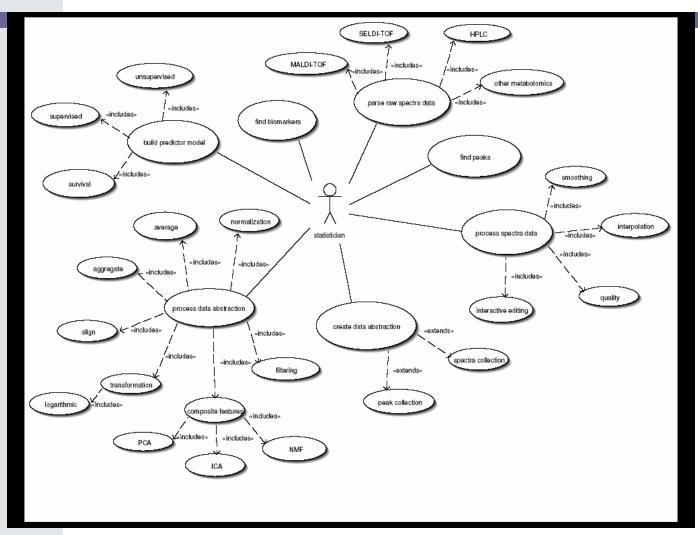
# **SMOS: UML model (part)**



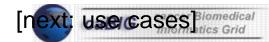




#### **SMOS: Use case Model**







# **Summary of Activities in the Proteomics SIG**

- ▶ LIMS Fox Chase
- ▶ Q5 Dartmouth
- ▶ Rproteomics Duke







# **Discussion**



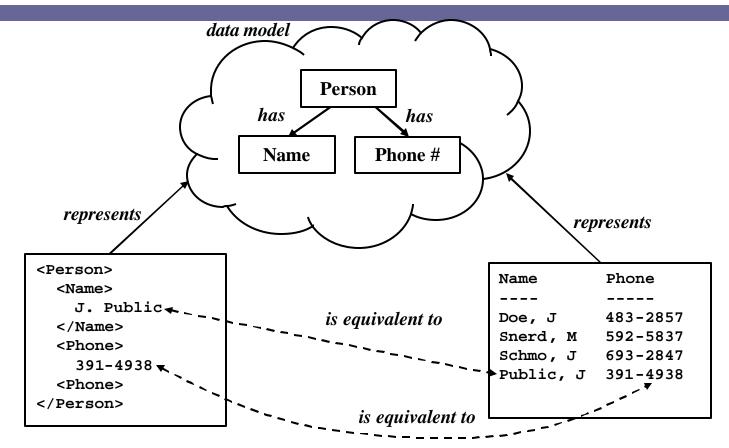
#### **Data and Metadata**

- What is data?
  - Observational (e.g., sensor readings)
  - Computational (e.g., output of simulation)
- What is metadata? data about data
  - Descriptive (e.g., biological specimen)
  - Relational (e.g., peak table generated from raw spectrum)
  - Contextual (e.g., units of measure)
- What's the difference?
  - Because metadata is still data, the difference could be blurred
  - Metadata is data that helps us use and understand other data





#### From Data Model to Exchange Format



interchange format 1: XML

interchange format 2: tabular





### Data Exchange

- Digital representation of data/metadata model (e.g., file, protocol message)
- Components of an interchange format
  - Syntax
    - Elements (e.g., area code, exchange)
    - Rules governing element types, occurrence, order, cardinality, etc. (e.g., area code is a three-digit integer which precedes the exchange)
  - Representation (e.g., XML, ASN.1, columnar)
  - Encoding (e.g., Unicode, ASCII, binary encoding)



